Supplementary File for Online Publication

For the paper:

Drug/Cell-line Browser (DCB): Interactive Canvas Visualization of Cancer Drug/Cell-Line Viability Assay Datasets

By:

Qiaonan Duan[1], Zichen Wang[1], Nicolas F. Fernandez[1], Andrew D. Rouillard[1], Christopher M. Tan[1], Cyril H. Benes[2], and Avi Ma'ayan[1,*]

[1]Department of Pharmacology and Systems Therapeutics, Systems Biology Center New York, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

[2]Center for Cancer Research, Massachusetts General Hospital Cancer Center, Harvard Medical School, 149 13th Street, Charlestown, Massachusetts 02129, USA.

Creating the cell-line canvases

There are four types of cell-line canvases, resulting in four different views in the DCB application. In these canvases, each tile represents a cancer cell line and the methods to make these canvases are described below.

Tissue of origin canvases

Each tile in the tissue of origin canvas is colored by tissue type. To construct the canvas, binary vectors are built for each cell-line indicating whether a cell line originates from a tissue. An adjacency matrix is computed from these vectors using the Jaccard index. The matrix is fed into the network2canvas algorithm [1] to generate the coordinates for each cancer cell line on the canvas in JSON format. DCB visualizes each cancer cell line as a SVG rect element using the D3 JavaScript library [2]. The cell-line and tissue (or subtype) mapping information was extracted from GDSC [3] Supplementary Data 1, CCLE [4] Supplementary Table S1 and Heiser et al [5] Supplementary Data 1.

Sensitivity scores canvases

Each tile in the sensitivity score canvas is colored by the clustering of cell-line's local fitness, which is the average distance to the cell's eight surrounding neighbors. The sensitivity profile of a cell-line is a vector of its sensitivity scores to all drugs. An adjacency matrix is computed from the sensitivity profiles using a customized Euclidean distance to deal with missing data:

$$distance = \sqrt{n/|M| \sum_{i \in M} (a_i - b_i)^2}$$

Where a_i and b_i are sensitivity scores in two response profiles **a** and **b**; n is the total number of cell lines; M is the set of cell lines that do not have missing values in both **a** and **b**. Similarly to the tissue of origin canvas, the matrix is fed into network2canvas algorithm and DCB visualize drugs as SVG elements on the drug canvas. The drug sensitivity data was extracted from GDSC Supplementary Data 1, CCLE Supplementary Table S11 and Heiser Supplementary Data 2. The sensitive scores used are IC50 for GDSC, Act Area for CCLE and GI50 for Heiser et al.

Gene expression similarity canvases

The methods listed below describe the process of generating the gene expression canvases.

GDSC Dataset: We downloaded the gene expression data for all cancer cell lines from the Genomics of Drug Sensitivity in Cancer (GDSC) project website at http://www.cancerrxgene.org/downloads/. This dataset consists of expression measurements of 13321 genes across 654 cancer cell lines. We performed quantile normalization after log2-transformation of the gene expression profiles and then converted the expression values for each gene to z-scores. We then obtained a 654 by 654 cell line similarity matrix by computing the Pearson correlation of the z-scores for every pair of cell lines. We set the values along the main diagonal to equal to zero and linearly transformed the off-diagonal similarity scores to range from 0 to 1. We expanded the matrix to include 60 cell lines with drug sensitivity data but with no gene expression data. We imputed similarity scores for these cell lines by setting all unknown values equal to the mean of the similarity scores among the 654 cell lines. We input the final 714 by 714 cell line similarity matrix into the Network2Canvas algorithm and obtained a canvas after ~6e8 attempted swaps (24 hours of run time on an Intel i5-3570 3.4 GHz quad core processor).

CCLE Dataset: We downloaded the gene expression data for the cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) website at http://www.broadinstitute.org/ccle/data. The data consists of log2-transformed expression measurements of 18901 genes across 1037 cancer cell lines. We performed quantile normalization on the gene expression profiles and then converted the expression values for each gene to z-scores. We obtained a 1037 by 1037 cell line similarity matrix by computing the Pearson correlation of the z-scores for every pair of cell lines. We set the values along the main diagonal to equal to zero and linearly transformed the off-diagonal similarity scores to range from 0 to 1. We then retained only the rows and columns of the matrix corresponding to the 493 cell lines for which CCLE has both gene expression and drug sensitivity data. We expanded the matrix to include 11 cell lines with drug sensitivity data but no gene expression data. We imputed similarity scores for these cell lines by setting all unknown values to equal to the mean of the similarity scores among the 493 cell lines. We input the final 504 by 504 cell line similarity matrix into the Network2Canvas algorithm and obtained a canvas after ~6e8 attempted swaps (24 hours of run time on an Intel i5-3570 3.4 GHz quad core processor).

Heiser et al. Dataset: We downloaded gene expression data for cancer cell lines from the publication's EMBL-EBI deposit at http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-181/, consisting of log2-transformed expression measurements of 23030 genes across 56 cancer cell lines. We performed quantile normalization on the gene expression profiles and then converted the expression values for each gene to z-scores. We obtained a 56 by 56 cell line similarity matrix by computing the Pearson correlation of the z-scores for every pair of cell lines. We set the values along the main diagonal equal to zero and linearly transformed the off-diagonal similarity scores to range from 0 to 1. We then retained only the rows and columns of the matrix corresponding to the 45 cell lines for which the Heiser et al study had drug sensitivity data. We input the final 45 by 45 cell line similarity matrix into the Network2Canvas algorithm and obtained a canvas after ~6e8 attempted swaps (24 hours of run time on an Intel i5-3570 3.4 GHz quad core processor).

Mutation Status Canvases

The mutation status canvases were tailored for each dataset. Hence, the processing of these datasets is described separately below.

GDSC Dataset: The mutation information of cell-lines was extracted from Supplementary Data 1. The mutation data includes copy number and point mutations of 64 commonly mutated cancer genes, MSI status and seven common gene translocations. We transformed the mutation data into a gmt file, where the terms are cell-lines and the sets are mutations in each cell-line. For the 64 commonly mutated cancer genes, we did not distinguish if the mutation is copy number variation or point mutation. As long as either type occurred, the gene is included in the gene set. Then, using the Sets2Network algorithm, we calculated a similarity adjacency matrix that was taken as input into the Network2Canvas algorithm to generate the canvas.

CCLE Dataset: Hybrid capture sequencing mutation dataset was downloaded from the Cancer Cell Line Encyclopedia website at http://www.broadinstitute.org/ccle/data in maf format. The file contains SNP and copy number mutation information for 904 cell-lines which include 509 out of the 639 cell-lines that are in the drug sensitivity dataset. We transformed the mutation data into a gmt file, where the terms are the 509 cell-lines and the gene sets are the mutated genes in each cell-line. Again, we did not distinguish if the mutation is copy number or SNP. As long as either type is present, the gene is included in the set. Then using the Sets2Network algorithm we calculated a similarity adjacency matrix that was taken as input to the Network2Canvas algorithm to generate the canvas.

Heiser et al Dataset: This study only includes information about seven commonly mutated cancer genes for each cancer cell-line. The data was extracted from Supplementary Data 1. We transformed the mutation data into a gmt file, where the terms are cell-lines and the sets are mutations in each cell-line. Then, the Sets2Network algorithm was applied to calculate a similarity adjacency matrix that was taken as input to the Network2Canvas algorithm to generate the canvas.

Creating the Drug Canvases

There are four types of drug canvases generated for each dataset, resulting in four different views in the DCB application. In these canvases, each tile represents a drug. The methods to generate these canvases are described below.

Sensitivity score canvases

Each tile is the sensitivity score canvas is colored by the drug’s clustering fitness measured as an average distance to its eight surrounding neighbors. The sensitivity profile of a drug is a vector of sensitivity scores of all cell lines in response to the specific drug. An adjacency matrix is computed from the sensitivity profiles using a customized Euclidean distance to deal with missing data as described above for the sensitivity score canvas created for the cell-lines. The drug sensitivity data was extracted from GDSC Supplementary Data 1, CCLE Supplementary Table S11 and Heiser et al Supplementary Data 2. The sensitivity scores used are IC50 for GDSC, Act Area for CCLE and GI50 for Heiser et al.

Chemical Structure Canvases

196 drugs in three datasets were mapped to canonical Simplified Molecular-Input Line-Entry System (SMILES) string via PubChem IDs, 2 of which (GSK2119563, GSK1487371) do not have PubChem records. Canonical SMILES strings of the 194 molecules were converted to 166-bit MACCS structural keys by the ChemmineR R package [6]. The drug structural similarity scores were computed using the Sets2Network algorithm based on the overlapping of the MACCS structural keys. Three drug-drug structural similarity networks consisting of 131 (GDSC), 24 (CCLE) and 74 (Heiser) drugs were created with the similarity scores computed as the weight of edges between drugs, the 2 drugs without structural information were set to have no edges with other drugs. The three drug-drug structural similarity networks were then fed into the Network2Canvas algorithm, with annealing time of 2, 10 and 10 hours, respectively. The drug information was extracted from GDSC Supplementary Data 2, CCLE Supplementary Table S6 and Heiser et al. Supplementary Data 1.

Target-based Canvases

We generated canvases of drugs based on similarity of drug targets from the drug-target information obtained from the GDSC Supplementary Data 2, CCLE Supplementary Data Table S6, and Heiser et al. Supplementary Data 4. Since drugs tend to share few targets, we calculated drug-drug similarity based on the distance of drug targets in literature-based protein-protein-interaction network (PPI) extracted from low throughput experiments. The construction of this network is described in our prior publication Expression2Kinases [7]. The drug-drug similarity measure was calculated as follows:

$$S_{DD} = \frac{1}{d_{ppi} + 1}$$

Where, S_{DD}, is the similarity measure between two drugs and, d_{ppi}, is the minimum distance between two drug targets in the PPI network. d_{ppi} was calculated using the shortest path length algorithm in the Python package NetworkX.

Drug Perturbation L1000 Canvases

We used the new data from the Library of Integrated Network-Based Cellular Signatures Connectivity Map (LINCS L1000 CMAP) to create canvases of drugs, where the drugs are clustered by similarity of the gene expression changes observed after drug treatment of cultured human cancer cell-lines. The LINCS L1000 data contains gene expression profiles following pharmacologic, genetic (over-expression or knockdown), or sham perturbation of cultured human cancer cells in a range of conditions, where a condition is defined by a particular cell type, time point, and compound concentration. This data can be obtained from http://lincscloud.org. Signatures of differentially expressed (DE) genes for a drug treatment can be obtained by comparing gene expression following drug treatment to gene expression following sham treatment in the same condition. A signature of DE genes generally takes the form of a vector of values matched to a set of genes, where the sign and magnitude of each value indicate the direction (up- or down-regulation) and significance of differential expression of each gene. Similarity of drugs can then be computed based on their signatures of DE genes.

The CCLE, GDSC, and Heiser et al studies profiled the sensitivity of cell lines to 24, 138, and 74 drugs, respectively. The information was extracted from GDSC Supplementary Data 2, CCLE Supplementary Table S6 and Heiser et al Supplementary Data 1. We searched for these drugs in the list of compounds used for the LINCS project and found 20 for CCLE, 110 for GDSC, and 52 for the Heiser et al. study. Next, we searched for a condition (cell line, time point, and drug concentration combination) with perturbation data for many of these drugs, so that we could compute signatures of DE genes for these drugs in the same condition. The condition with the most data for these drugs was the MCF7 cell line with drug treatment of 10 µM for 6 hours, which had gene expression data for 19 of the CCLE drugs, 106 of the GDSC drugs, and 47 of the Heiser et al drugs. Gene expression profiles following perturbation with these drugs were measured in multiple replicate cell culture experiments. For each drug perturbation experiment, we used the Characteristic Direction [8] to compute a signature of DE genes relative to plate-matched sham perturbations. We then averaged replicate experiments to obtain a single signature of DE genes for each drug in the MCF7, 10 µM, and 6 hour condition.

References

1. Tan CM, Chen EY, Dannenfelser R, Clark NR, Ma'ayan A: **Network2Canvas: network visualization on a canvas with enrichment analysis**. *Bioinformatics (Oxford, England)* 2013:btt319.
2. Bostock M, Ogievetsky V, Heer J: **D³ data-driven documents**. *Visualization and Computer Graphics, IEEE Transactions on* 2011, **17**(12):2301-2309.
3. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P, Thompson IR, Luo X, Soares J: **Systematic identification of genomic markers of drug sensitivity in cancer cells**. *Nature* 2012, **483**(7391):570-575.
4. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, Sonkin D: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity**. *Nature* 2012, **483**(7391):603-607.
5. Heiser LM, Sadanandam A, Kuo W-L, Benz SC, Goldstein TC, Ng S, Gibb WJ, Wang NJ, Ziyad S, Tong F: **Subtype and pathway specific responses to anticancer compounds in breast cancer**. *Proceedings of the National Academy of Sciences* 2012, **109**(8):2724-2729.
6. Cao Y, Charisi A, Cheng L-C, Jiang T, Girke T: **ChemmineR: a compound mining framework for R**. *Bioinformatics (Oxford, England)* 2008, **24**(15):1733-1734.
7. Chen EY, Xu H, Gordonov S, Lim MP, Perkins MH, Ma'ayan A: **Expression2Kinases: mRNA profiling linked to multiple upstream regulatory layers**. *Bioinformatics (Oxford, England)* 2012, **28**(1):105-111.
8. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, Ma’ayan A: **The characteristic direction: a geometrical approach to identify differentially expressed genes**. *BMC bioinformatics* 2014, **15**(1):79.